Chapter 4 – Sampling Design

INTRODUCTION

The NPS Inventory and Monitoring Program and other investigators with experience designing comprehensive and multidisciplinary monitoring efforts (e.g., Schreuder et al. 2004, Ringold et al. 2003) argue that individual protocols should be linked spatially, ecologically, and statistically. In an effort to integrate protocols, we began development of 10 protocols (Table 4.1) with a multidisciplinary team of university, USGS, and NPS scientists. We elected to begin with these 10 protocols, which cover 17 Vital Signs, and: a) span the types of habitats that we expect to monitor (e.g., aquatic, terrestrial, and airborne); b) demonstrate clear ecological linkages and high potential for data integration; and c) force us to consider sampling design at several levels of ecological organization and spatial resolution (e.g., landscape, communities, and species; Table 4.1). These protocols will be implemented over the first two years of the program (2006 and 2007), and additional protocols will be added in the ensuing years (see Chapters 5 and 9).

In this chapter, we present overviews of our efforts to develop an initial set of protocols and provide brief summaries of four of them. We discuss types and components of monitoring designs, underlying concepts, and justifications for the designs we have chosen. We also discuss integration among the protocols and with other monitoring efforts. Definitions of some terms used in this chapter are provided in the glossary (Appendix C) and in Box 4.1; these terms appear in bold upon first use in the text.

Inference-based and Non-random Designs

Monitoring programs should be based on statistically robust sampling designs when possible and should be broadly accepted by the scientific community (Christensen et al. 1996). Most of our protocols will be 'inference-based' so that data can be used to describe the entire park or large portions of it. However, it is sometimes necessary to adopt data collected by others, even if those data are collected at sites that were located in a non-random fashion, or if there was low sampling intensity. For example, NOAA weather stations, EPA air-quality stations, and USGS stream gages are not located randomly, and often too few stations exist to provide the statistical power needed to detect change over time within any given park. Yet, the Network is not able to invest adequate funding to improve substantially on these efforts and we will therefore use these available data to monitor some aspects of ecosystems. Similarly, parks in the Network have collected important data over the years for some Vital Signs and it is desirable to adopt and build on these efforts to maintain continuity. In other cases, management questions dictate that we sample at specific areas such that an inference-based approach is not appropriate. We use the term 'non-random' to describe the Network's use of such directed sampling, use of existing partner data, and adaptation of existing monitoring protocols. Consequently, Vital Signs within the Great Lakes Network will be sampled according to one of the two design types described below.

Inference-based Designs

Most Vital Signs will be monitored under protocols that we write according to NPS standards (Oakley et al. 2003) where we will select sample points **probabilistically** to maximize our ability to make inferences to a larger population. These designs will have a high level of statistical rigor and we will ensure adequate sample intensity by conducting simulation or **power analyses** based on comparable or past data sets. Inference-based designs include those for water quality of large rivers, terrestrial vegetation, and amphibians as well as other Vital Signs in the future.

Non-random Designs

Some Vital Signs will be monitored under protocols that we write according to NPS guidelines (Oakley et al. 2003), but from which we will be unable to make statistical inference to a broader area. These are cases where sampling design is predetermined or substantially modified by existing monitoring efforts or where management questions indicate that we direct our sampling to specified areas. By adopting past methods, even when inferences can not be made to a broader area, we can maintain historical and regional datasets that provide spatial and temporal context for the parks. This includes, for example, maintenance of landbird monitoring data that have been collected in a similar fashion for many years across the Great Lakes region by several agencies. By making slight modifications to these existing protocols and clearly documenting the procedures, we will increase consistency and repeatability. Similarly, we will use a non-random design to monitor specific areas or resources (e.g., a set of lakes) when it is not feasible to sample randomly or desirable to make inference to other areas. In all cases, we will examine the quality and completeness of past data, conduct simulation or power analyses to assess the adequacy of sample size, specify (i.e., qualify) the **sampling domain**, consider improvements of the domain, make the data available for analyses of other Vital Signs, and periodically summarize them. Non-random designs include those for air quality and landbird data that have been collected by parks and partners, weather data from NOAA, and stream gage data from USGS.

Box 4.1. Terms used in Chapter 4 (see also the Glossary in Appendix C).

Alpha (α) – The predetermined threshold of statistical significance in null-hypothesis testing. This threshold is frequently set at 0.01, 0.05, or 0.1. *P*-values less than alpha suggest a phenomenon that would rarely occur by chance alone (e.g., a strong trend, relationship between variables, or difference between groups); tests with *P*-values greater than alpha are deemed 'non-significant.'

a priori – Beforehand; when referring to power analyses, this refers to analyses conducted prior to sampling that use existing data to obtain estimates of variability in the monitored component to either estimate sample sizes needed to detect a desired level of change or determine what amount of change can be detected with a particular sample size (see 'Power,' below).

GRTS – Generalized random tessellation stratified (GRTS) design strategy. This design allocates samples in a spatially balanced manner to either linear systems (e.g., a stream network) or other sampling areas (e.g., forest patches). Also maintains spatial balance with addition or deletion of samples.

Power – The probability that a test will reject a false null hypothesis, or in other words that it will not make a Type II error. Power increases as sample size or effect size (e.g., magnitude of change) increases, variability in the indicator decreases, and as alpha is relaxed (= increased).

Power analysis – A calculation performed to estimate sample sizes needed to detect a desired level of change or determine what amount of change can be detected with a particular sample size. Power is a function of sample size, sample variance, effect size, and alpha; consequently, if any four of these variables are known (or chosen), the fifth can be calculated.

Probabilistic design – A sampling design in which all potential points within the sampling domain have a known probability of being selected for sampling. Selection occurs via some process that randomly selects points.

Sample panel – A group of sample units visited at the same recurring interval. Sampling units (e.g., sites) from the entire population may be subdivided into several panels, each of which may be sampled more or less frequently, depending on the re-visit strategy.

Sampling domain – The area in which samples occur. If sampling locations are randomly selected and have reasonable replication, this corresponds to the area about which inferences can be drawn.

Simple random sampling -- strategy in which the number of total sampling sites is selected from the sampling frame (i.e., domain of interest), such that every point within the target area has the same probability of being selected. The procedure for selecting units must be truly random.

Stratified random sampling – sampling strategy in which the overall domain of interest (i.e., sampling frame) is divided up into mutually exclusive and exhaustive subpopulations called strata, each of which is clearly defined. Each sampling unit is subsequently classified into the appropriate stratum, and then a simple random sample is drawn from each stratum.

Systematic sampling – a sampling algorithm in which the first sampling unit is randomly selected and subsequent units are selected according to a regular (i.e., systematic) pattern (e.g., every *i*th grid cell) (Mendenhall et al. 1971)

Type I error –Incorrect rejection of a null hypothesis that is actually true. For example, it is stated that a trend is detected when, in fact, none exists. When expressed as a probability, it can be symbolized by alpha (α); when expressed as a percentage, it is known as significance level.

Type II error – Failure to reject a false null hypothesis. For example, concluding that no trend (or no trend of a particular magnitude) has occurred, although one actually has.

Table 4.1. Habitats, ecological attributes, and linkages of Vital Signs that will be monitored as part of an initial set of protocols being developed by the Great Lakes Inventory and Monitoring Network.

Protocol	Vital signs being covered	Habitat	Ecological attribute	Ecological linkages between protocols
Air Quality	Air Quality	Air	Chemical and process	Major driver of change affecting each of the other indicators; air quality impacts water quality through wet and dry deposition
Climate and Weather	Weather	Air	Process	Major driver of change that affects each of the other indicators
Land Cover / Land Use Coarse Scale	Land Use Coarse Scale	Aquatic and terrestrial	Landscape	Major driver of each of the other indicators; e.g., land cover affects water runoff, quality of water and air, health of many vertebrate species
Land Cover / Land Use Fine Scale	Land Use Fine Scale, Stream Dynamics	Aquatic and terrestrial	Landscape	Major driver of each of the other indicators; e.g., land cover affects water runoff, quality of water and air, health of many vertebrate species
Terrestrial Vegetation	Terrestrial Plants, Succession, Problem Species (in part), Terrestrial Pests and Pathogens, Soils	Terrestrial	Species, community, and process	Affected by weather patterns, land use, and air quality; potential buffer for water quality; habitat for landbirds
Water Quality for Inland Lakes	Core Water Quality Suite, Advanced Water Quality Suite, Water Levels	Aquatic	Chemical and process	Affected by weather patterns, land use, and air quality; affects amphibians, diatoms, fish, and bioaccumulation of toxics
Water Quality for Large Rivers	Core Water Quality Suite, Advanced Water Quality Suite, Water Flow	Aquatic	Chemical and process	Affected by weather patterns, land use, and air quality; affects amphibians, fish, benthic invertebrates, and bioaccumulation of toxics
Amphibians	Amphibians and Reptiles (in part)	Aquatic and terrestrial	Species and community	Indicators of water quality; may also reflect changes in climate, land use, and land cover; are consumed by birds and other predators
Bioaccumulative Contaminants	Trophic Bioaccumulation; Species Health, Growth and Reproductive Success	Air and aquatic	Process and species	Assess the ecological effects of air- and water-borne toxics that biomagnify in the environment
Landbirds	Bird Communities	Terrestrial	Species and community	Affected by patterns and magnitude of land use, terrestrial vegetation, and climate

DESIGN COMPONENTS AND CONCEPTS

Sampling Domains

One of the essential components of a sampling design is a clear identification of the sampling domain (i.e., the area effectively sampled), including a precise description of the target population. The 'target population' is the ecological resource for which information is desired. The population may be discrete, as in the population of lakes within a park boundary, or continuous, as in a tract of forest land or a length of stream. We used an iterative process that included conceptual models and meetings with park and partner scientists to develop monitoring questions, which, in turn, identified target populations and sampling domains.

The nature of the target population guides the development of a sample design. If the target population is small enough that it can be sampled in its entirety (i.e., a census approach), then statistical inference is not an issue. More often, though, the target population will be large relative to our sampling capabilities, and a representative sample must be selected. Ensuring that a sample is truly representative of the target population is a key consideration in development of GLKN protocols, but this consideration must be balanced against logistics, safety, and cost (Field et al. 2005).

Park boundaries pose a significant challenge to monitoring programs because the stresses imposed on park resources often originate outside of park boundaries. While physical sampling outside the park boundary is often not possible or economically justifiable, the Network will use remotely sensed data to assess changes in land cover and land use not only within park boundaries but also in buffer areas around each park.

Spatial and Temporal Allocation of Samples

Given a large target population, the sampling designs least likely to produce bias are those in which samples are selected probabilistically (Manly 2001, Hayek and Buzas 1997). McDonald (2003) provides terminology to discriminate between the spatial and temporal components of a survey design.The *membership design* describes how sample units are selected spatially, and the *revisit design* describes how often individual units are sampled over time. Many alternative membership designs were considered in the GLKN effort, including simple random, stratified random, and systematic sampling, as well as designs that more strongly accommodate logistical and safety constraints. One design that we have used and plan to use in other, future protocols is the generalized random tessellation stratified (GRTS) design strategy (Stevens and Olsen 2004, 2003). This design allocates samples in a spatially balanced manner to either linear systems (e.g., a stream network) or desired sampling areas (e.g., forest patches on an archipelago or in a Lakeshore). The design allows for iterative addition or deletion of samples, while maintaining spatial balance at several hierarchical spatial scales. Several designs were discarded because of inherent disadvantages (e.g., see Table 4.1 of Jean et al. 2004). For example, when total sample size is small relative to the area sampled, simple random sampling may result in samples that are overly clustered, and by chance alone may mean that certain regions of the target population are not sampled. Stratified random samples have the advantages of increased efficiency and precision, but require that the strata be

delineated accurately and persist over time (Stevens and Olsen 1991; D. Stevens, Oregon State University, personal communication).

The revisit design was also a critical consideration for our protocol development (Table 4.2). The choice of revisit design involves tradeoffs among the ability to detect interannual trends, the ability to describe spatial variation in a response variable, and the cost of collecting each sample.

The actual designs used for most of our protocols are one of two variants. In *repeating panel designs*, groups of sample units, or **sample panels**, are revisited at a recurring interval. For example, all river sites at SACN comprise a panel, which will be sampled every other year. We may also be using *split-panel designs* (using two or more revisit designs; McDonald 2003); for example a subset of inland lakes will be sampled for water quality every year at each park, and the remaining lakes may be sampled on a longer rotation (e.g. every 10 years).

In the final analysis, accessibility, sampling cost and safety became critical constraints that were factored into the development of designs for several protocols. Additionally, GLKN staff and park personnel recognized a number of instances where it was important to maintain or create **'index' sites** – sites selected for sampling because they are of particular interest, or because they have a legacy of long-term sampling (which allows us to conduct retrospective analyses). Because the area represented by such index sites is difficult to quantify, index sites will not be combined with probabilistically selected sites in statistical analyses.

Sampling Intensity and Frequency

In general, sample size should be large enough to give a high probability of detecting any changes that are of management or conservation importance, but not unnecessarily large (Fry 1992). To estimate appropriate sample sizes, we performed (or will perform) ***a priori*** power analyses, simulation modeling, or both. *A priori* power analyses are statistical calculations made prior to the initiation of monitoring fieldwork using pre-existing data (Thomas and Krebs 1997). Because these data provide an estimate of the variability in the target indicator, power analyses can be used to estimate the approximate sample size needed to detect a trend of a given magnitude. For power analyses, we used 20% as a minimum level of change that we sought to detect. Most resource managers at our parks felt this detection level was reasonable, and other monitoring programs have adopted this standard as well. We were interested in detecting change in either direction (i.e., whether it were an increase or decrease in the indicator); we thus employed two-tailed tests. We used web-based power calculators and simulation analyses to determine how many sampling locations the Network would need to detect a 20% change between two points in time, in a paired *t*-test framework. In these analyses, the period of time over which the change occurs is not inherently specified. Instead, the temporal period depends on how many years occur between sampling events.

Table 4.2. Monitoring approach for ten protocols being developed by the Great Lakes Inventory and Monitoring Network in 2006 and 2007.

Protocol[1]	Sampling approach	Spatial sampling design	Revisit design and sampling frequency	Domain of inference
Air Quality	Acquire park and partner data	Index sites; stations in and adjacent to each park	No panels; all stations engaged in continuous data collection	Stations will only index interannual change at each site; kriging or field sampling may be used to interpolate to other park areas
Weather and Climate	Acquire park and partner data	Index sites; stations in and adjacent to each park	No panels; all stations engaged in continuous data collection	Stations will only index interannual change at each site; kriging or field sampling may be used to interpolate to other park areas
Land Cover / Land Use Coarse Scale	Satellite imagery	Entire park with larger regional extent for context	complete revisit every 5-7 years	Entire park area, and adjacent areas (watersheds or 10 km buffer)
Land Cover / Land Use Fine Scale	Aerial photography	Entire park with adjacent buffer	Complete revisit every 5-7 years	Entire park and 400 m to 2 km buffer depending on park
Terrestrial Vegetation	Site visits with plots and transects	Grid-based GRTS plus index sites	Entire park, complete revisit every 5 years	Entire park area that is forested, except some smaller islands at ISRO, VOYA and APIS
Water Quality for Inland Lakes	Site visits and acquire partner data	Index sites	Complete revisit, annually, 3x/yr	Individual lakes
Water Quality for Large Rivers	Site visits and acquire partner data	Linear-based GRTS and index sites	Complete revisit, every other year, monthly during open-water season	Mixed, due to use of both randomly selected and index sites
Amphibians	Site visits along roads, or fixed-area searches	Simple random; grid-based and linear GRTS	Ideally, complete revisit, annually. Still being debated.	Pilot work will determine whether road-based or entire park
Bioaccumulative Contaminants	Site visits to sample individuals	Census of nests or colonies; census or random sample of tissue for lab analyses	Complete revisit or repeating-panel; annual to every 2-3 years	Buffers around individual nests (eagles), individual-based areas for other species
Landbirds	Acquire park-collected off-road point data	Points placed systematically along transects	Complete revisit, annually	Historic designs placed transects haphazardly (non-randomly), and thus produce only an index

[1] = See Table 4.1 for a list of Vital Signs being monitored under each protocol.

In addition, to determine how many consecutive sampling events (across years) would be required to detect a 20% change in water-quality variables at each lake in the network, we used analyses (Gerrodette 1993) of root mean-square error using historical data. We are not aware of currently available power analyses that simultaneously incorporate spatial, intra-annual, and interannual variability; one can ask either how many sampling locations are needed, or how many repeat years of sampling are needed to detect a selected level of change.

For complex monitoring designs that may need to account for issues such as detection probability, fixed and random effects, and missing data, simulation modeling can be a particularly useful approach for determining sample size (Eng 2004, Muthén and Muthén 2002, Lukacs, *in prep*.). Simulation modeling employs a mathematical model to virtually repeat the study hundreds or thousands of times, to allow estimation of power essentially by direct measurement (Eng 2004).

Type I versus Type II Errors

As with all scientific hypothesis testing, monitoring programs must weigh the relative costs and benefits of **Type I** versus **Type II errors**, and set **alpha** (α) and **power** ($1 - \beta$) accordingly (Field et al. 2005, Di Stefano 2001, Steidl et al. 1997, Toft and Shea 1983). Scientists traditionally seek to reduce Type I errors and accordingly prefer small alpha levels (Shrader-Frechette and McCoy 1992). In a monitoring program with a strong resource-conservation mandate, however, it may be preferable to employ an early-warning philosophy by increasing alpha and consequently increasing the power to detect differences or trends (Roback and Askins 2005, Sokal and Rohlf 1995, Shrader-Frechette and McCoy 1992).

Accordingly, we have adopted an alpha = 0.10 and power = 0.80, to be able to detect magnitudes of change of $\geq 20\%$, in agreement with other NPS I&M approaches. Furthermore, we recognize that analyses investigating resource degradation whose results involve $0.20 > \alpha > 0.10$ may merit increased monitoring or experimental research.

For our initial set of protocols, *a priori* power analyses were conducted when possible to determine the approximate sample size needed to detect meaningful ($\geq 20\%$) levels of change. Given our specification of alpha, desired power, and effect size, combined with information on the variance of the response variable in question (obtained from past or comparable monitoring), it was possible to calculate the sample size required to achieve these results. In some cases it was necessary to abandon measurements of highly variable indicators or qualify the resulting data as being useful only for showing the range of variability.

In several instances the program TRENDS (Gerrodette 1993, 1987) was used to perform power analyses to estimate sample sizes. One key decision in any power analysis involves determining the estimate of variance. When assessing power to detect trend across a spatial domain, the coefficient of variation among sampling locations has traditionally been used. Most of the parks, however, are interested in detecting interannual trends in Vital Signs. We acknowledge that TRENDS and most other power analysis programs can handle only very simple designs, and will not give a true

indication of power when revisit designs and measurement panels become more complicated. These programs were therefore used as heuristic rather than exact methods for estimating power, by providing a first approximation of required sample sizes. We will use simulation approaches to generate a more accurate estimate of power once an initial data set is obtained.

For analysis of temporal change at a single sampling location, it is more appropriate to use the Root Mean Square (RMS) error derived from a linear regression of response-variable data over time – essentially the coefficient of variation around the regression line (Nur et al. 1999). The RMS has the advantage of addressing an important component of variation – the scatter around the prediction line when a trend is present – and incorporates numerous sources of error, including random measurement error, sampling error, and the inherent variation around an individual observation. With respect to trend analyses, this analysis yields the number of repeat sampling events (i.e., across, not within) years required to detect a significant trend at that sampling location.

RESULTING DESIGNS

For each protocol, we adopted sampling designs that best met the following considerations: ability to answer the monitoring question(s), applicability to the domain(s) of interest, conformity to standards of the discipline, statistical power, comparability of data to regional or national monitoring programs, suitability for retrospective analyses (i.e., ability to incorporate pre-existing, longer-term data), logistical constraints (accessibility), safety, and cost. Each protocol, and often each park, had unique problems and thus no one design fit all applications. The following sections describe key design aspects of four protocols that were pilot-tested in 2006 and will be further tested or ready for implementation in 2007. Protocol Development Summaries are available for these four in Supplemental Document 7. Several other protocols are also under development, and are summarized in Supplemental Document 7, but their sampling designs have not been fully addressed. We envision that the proportion of protocols that utilize probabilistic sampling will continue to increase over time, although in some parks the spatial domain may be limited (e.g., for especially inaccessible or unsafe sites).

Water Quality for Large Rivers

Sample design for the large rivers protocol was derived in part from two established ecological monitoring efforts, the USGS National Water Quality Assessment program (NAWQA) and the EPA's Environmental Monitoring and Assessment program (EMAP). The NAWQA program uses two types of fixed sites: integrator sites, which are located at major confluences of tributaries with the mainstem, and indicator sites, which are believed to represent conditions in relatively homogeneous basins. The EMAP program uses a generalized random-tessellation stratified (GRTS) design that results in a spatially dispersed yet random sample (Stevens and Olsen 2004). The sampling design for GLKN rivers differs from the EMAP and NAWQA approaches in that it uses a combination of randomly-selected and index sites. Selection of random sites involves a GRTS approach, by distributing a target number of sites (derived from power analyses) along the length of the mainstem of the river (Figure 4.1). This approach will be applied across the St. Croix and Namekagon Rivers, within SACN. Power analysis on past data has shown that six randomly selected sites, three each in the upper and lower portions of

the riverway, are adequate to meet our criteria for detecting interannual change in most water-quality variables. For separate analyses, index sites will be selected based on recommendations from the multi-agency St. Croix Basin Water Resource Planning Team, which is currently developing a comprehensive monitoring plan for the basin. Based largely on budget considerations, we expect to select five integrator sites along the St. Croix and Namekagon Rivers. Many sites on the Mississippi River within MISS park boundaries are currently monitored by other agencies. We have selected additional index sites at MISS to fill gaps where stretches of the river are not included in monitoring conducted by others.

The randomly selected sites in SACN will allow inference across the mainstem of the entire St. Croix and Namekagon Rivers, within park boundaries. The integrator sites will not allow inference to other areas of the rivers, and data from these sites will be analyzed separately for each site, through time. At MISS, sampling sites were selected to add information to ongoing monitoring programs. Thus, data will again be analyzed separately for each site.

Sampling will alternate yearly between the two large river parks. During each sampling year, the rivers will be sampled nine times during the open water season, approximately monthly, from May to November.

Lotic systems such as large rivers provide a potential challenge, in that the same water that exists at a given point in time will occupy a point downstream at a later point, albeit after mixing, dilution, and dispersion. Hydrologists acknowledge that downstream locations are thus partially dependent upon upstream locations, although upstream locations are not influenced by what happens to water quality downstream of them. However, they also recognize that characteristics of a sampling point's drainage area (i.e., its geology, geomorphology, land use, etc.) will influence the water quality at that sampling point. That is, if the water quality at a downstream location is different from an upstream location, we attribute those water quality differences to the intervening drainage area. As long as the time and distance between the two samples exceeds the residence time or flow rate of the river, then a hydrologist usually expects the samples to be independent of each other. In our work, we are fairly confident that the study design employs independent sampling locations. At SACN, 11 stations are spread over a large drainage area (nearly 7800 sq. mi.), spaced by meaningful distances, such that we expect water-quality results to be independent among stations. Furthermore, the most closely located stations in the design, the three random sites in Lake St. Croix, are located in 3 of 4 separate sub-basins of the lake, and will be sampled in a downstream-to-upstream sequence (reducing the possibility of "replicate" water-quality results).

Water Quality for Inland Lakes

Great Lakes Network parks contain hundreds of inland lakes, with 299 occurring at VOYA alone. In our first attempt to design an inland lake monitoring protocol, we limited our domain of interest by lake size, depth, and accessibility. We defined lakes as waterbodies with a surface area > 1 ha and a maximum depth of > 1 m, to be consistent with definitions used by the federal EPA-EMAP program and others in states of the upper Great Lakes. We also limited our domain to lakes that are accessible via road or trail because many of the lakes at VOYA and ISRO would require two or more days of off-

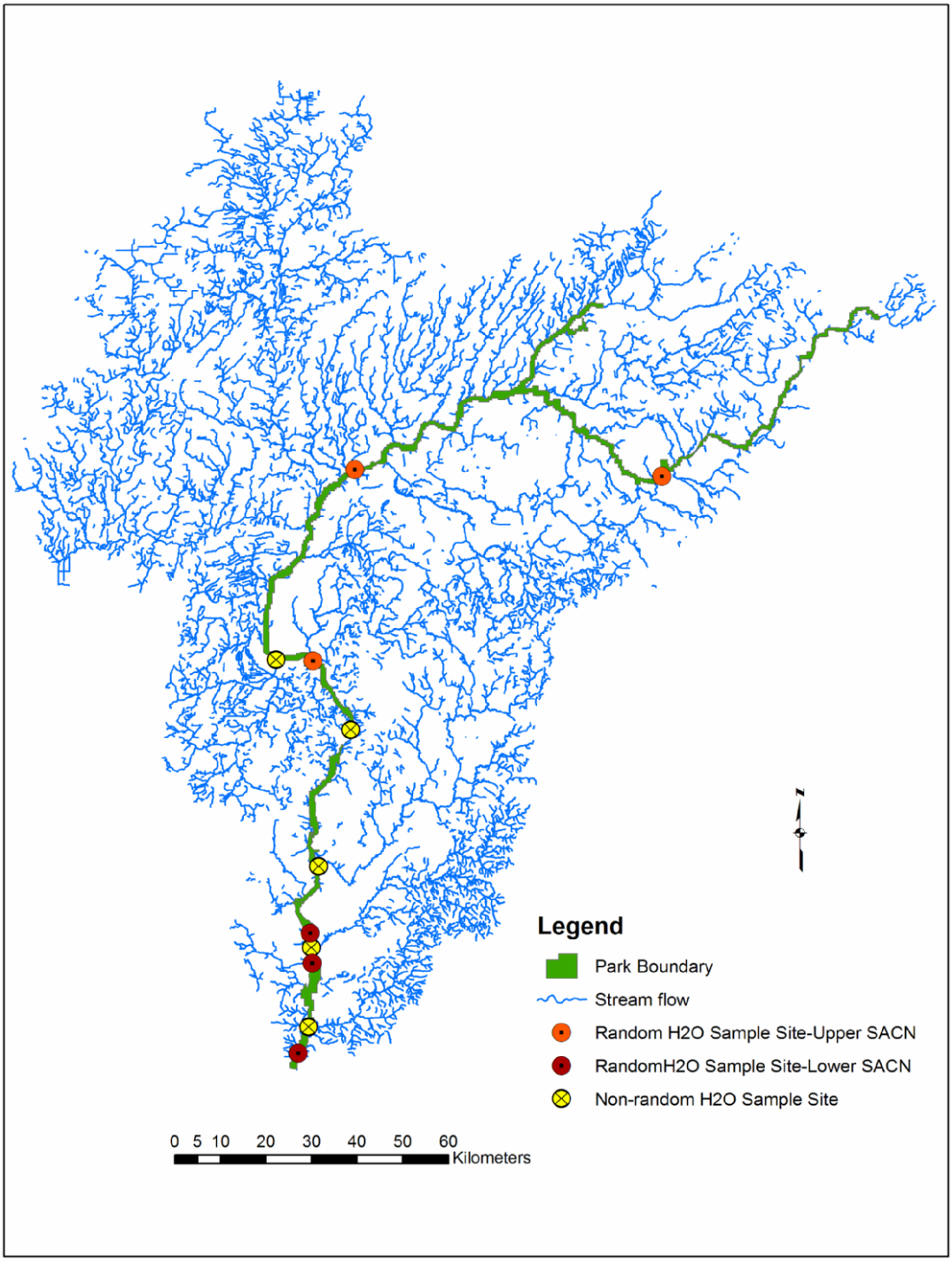

Figure 4.1. Location of randomly selected and index sites for monitoring water quality on the St. Croix and Namekagon Rivers, SACN. Blue lines depict tributaries.

trail back country travel to access. Compounding the access constraints is the need to maintain water samples in cold, dark conditions prior to analysis. Our design resulted in a census of lakes within our defined domain of interest, with all lakes of interest being sampled at some point within the revisit design. Most lakes would have been sampled on a 3-year rotation, with some lakes being sampled on a much longer rotation (e.g. 9 years at VOYA). This approach would not have allowed extrapolation of monitoring results to unsampled lakes.

The most consistent and substantial criticism we received during the peer-review process was in regard to the revisit frequency. Limnologists pointed out that a 3-year rotation could coincide with other cycles, such as El Nino or years of strong fish age classes, and that the amount of time it would take to detect potential trends (27 years in the cases of those lakes sampled on a 9 year rotation) was too great. We are thus revising our design to sample fewer lakes every year.

We are currently working with parks to select lakes within the same size, depth, and accessibility constraints as above. It is likely that the lakes will not be selected randomly, but rather will be selected based on management concerns and amount of historic data. We will strive to select lakes that are spatially dispersed within each park and span a gradient of current water quality conditions and levels of recreational use. When information exists on types of lakes within a park, such as that by Carlisle (2002) for ISRO and Schupp (1992) for VOYA, we will attempt to select lakes from each category.

The frequency of sampling within a year, sample locations, and parameters sampled are designed to allow integration and comparisons with data collected by state and other agencies. The nonrandom selection of lakes in our design, however, will not allow for inferences to lakes other than those sampled. We will analyze data from each lake separately and will use correlational statistics to determine whether parallel trends occur among lakes within a park, across parks, and within the larger region. When similar trends are observed in multiple lakes, additional monitoring may be warranted to determine whether the trend is ubiquitous. Research may also be warranted to determine the cause of the trend.

Amphibians

To be comparable with long-standing amphibian monitoring programs, such as the North American Amphibian Monitoring Program (NAAMP; Weir 2005), Marsh Monitoring Program (MMP; Timmermans et al. 2004), and Amphibian Research and Monitoring Initiative (ARMI), our design will incorporate aspects of each. Our draft protocol recommends a combination of nighttime call surveys (at GRPO, INDU, MISS, PIRO, SACN, SLBE) and daytime visual encounter surveys (at APIS, ISRO, and VOYA).

During the first two years, 2006 and 2007, we will conduct intensive monitoring at a subset of sites at three parks to model detectability for estimating site occupancy (MacKenzie et al. 2004, 2003, 2002) for each species we expect to encounter. In this pilot work, we will also test the effectiveness of parabolic reflector microphones and remote call-recording devices in monitoring and recording calls beyond road corridors to include more remote areas of the parks. In 2008, we will make revisions to the draft protocol with

the potential for broadening the monitoring to include more sites and all or a subset of the nine parks.

The sampling design(s) chosen for the nighttime call surveys will depend upon the effectiveness of the parabolic microphones and recording devices, as well upon whether park managers prefer inference to the whole park area or prefer greater sample sizes (and thus, greater precision) at the expense of a reduced (and perhaps biased) sampling domain. Because the initial plan involves limiting nighttime call surveys to roads, the area of inference for the nighttime call surveys will be limited to a buffer around roads equal to the maximum distance at which species can effectively be detected. We recognize, however, that surveys conducted along roads are inherently biased because: a) the roads themselves are not randomly located (i.e., they are often routed around the wetland habitats preferred by many amphibians); b) road-associated stressors (e.g., road salts, noise and dust generated by vehicular traffic, discarded trash, vectors of non-native species) disproportionately affect wetlands at different distances from roads; and c) the road geometry itself creates unequal probabilities of including different sites (e.g., a site might be accessible from portions of two different roads). We will need accurate wetland maps to calculate the probabilities of inclusion (D. Stevens, Oregon State University, personal communication). When this protocol is fully implemented, observers for nighttime surveys will identify up to eleven frog and toad species at up to 30 randomly chosen sites, although we may be unable to select 30 sites in GRPO and APIS. For daytime surveys, the list could include two salamander species as well. In our initial years at each park (during which detectability must be modeled, to correctly understand and interpret trends), sites will be visited three times during each of three sampling periods per year. For broader-scale-analyses (e.g., across the Network), pooling of sites can only occur when sampling with the same method; thus, daytime and night-time sites will be analyzed independently.

Due to the lack of roads at VOYA, ISRO, and APIS, we will conduct daytime surveys using a combination of call surveys, dip-net sweeps, and wetland perimeter searches. The sampling domain will be limited to wetlands within 1000-m buffer areas along the shoreline of Lake Superior, other large lakes, roads, and trails. Defining our domain in this way will allow a reasonably large proportion of these three parks to be sampled. The sampling areas (i.e., park units) will be divided into 6.25-ha (15.4-ac) cells using the GRTS method (Stevens and Olsen 2004). From this initial set, the first 30 cells that contain habitat for wetland-breeding amphibians will be sampled annually. We will use percent area occupied (PAO) as the primary metric and build models of detectability over the first two to three years of the effort. Additional data will include numbers caught or observed within each age class (eggs, metamorphs, adults) per unit effort. Revisit strategies are still being debated (e.g., we will convene an amphibian expert panel in Feb. 2007), though the great interannual variability in amphibian populations (especially in population size, but also in occupancy; L. Bailey, *unpubl. data*) argues for sampling every year.

For both nighttime and daytime amphibian surveys, environmental data such as weather and water quantity and quality are collected as covariates, for use in comparing various models that describe heterogeneity in occupancy.

Bioaccumulative Contaminants

This protocol is designed to monitor concentrations of bioaccumulative contaminants in tissue samples from bald eagles, herring gulls, and one additional species (under development) that inhabit aquatic systems of parks in the Great Lakes Network. The species, and thus strategies for monitoring, will depend on the species' abundance and distributions within each park. We will target legacy and emerging contaminants that are of concern to human and ecosystem health including mercury, lead, PCB's (polychlorinated biphenyls) and DDT (dichlorodiphenyltrichloroethane).

Bald eagle nestlings will be sampled from all known active nests in APIS, MISS, PIRO, SLBE, SACN, VOYA and ISRO by taking up to 11 cm^3 of blood and by plucking four feathers from each nestling. This effort relies on a significant partnership with Clemson University, which is collecting all of the data for parks in Michigan (SLBE, PIRO, and ISRO) and for one park in Minnesota (VOYA; as a control), as part of the Michigan Department of Environmental Quality's Wildlife Contaminant Trend Monitoring Program (Roe et al. 2004). The GLKN will take responsibility for collecting contaminants data from bald eagles at the remaining Network parks that have adequate numbers of eagles (APIS, MISS, and SACN). Both Clemson and GLKN will attempt to gather samples from all active nests in each park (i.e., perform a census). However, the proportion of tissue samples analyzed for contaminants in a given year will depend on per-sample costs, variability in concentrations of the various contaminants, and the number of active nests in each park.

During pilot work in 2006 the GLKN team sampled bald eagle nestlings from 32 of the 37 nests that were known to be active in APIS (n = 8 nests), MISS (n = 10), and SACN (n = 14). Up to two nestlings were captured opportunistically at each nest (i.e., the first and second nestlings that could be most readily captured). We could not sample from five nests because the young were too old to handle safely. Laboratory analysis will be completed on tissue samples from the nestling with the most complete sample (e.g. 11 cm^3 of blood and four feathers), because a full 11 cm^3 of blood is needed for analysis of all analytes. We may not be able to afford laboratory analysis on all samples in future years. If we must limit the number of samples analyzed for contaminants, we will do so using either a simple random, stratified–random, or spatially balanced design. We will stratify only if there is reason to believe that a spatial gradient exists for the contaminants being monitored. Non-analyzed samples will be archived for future use. Previous work by Roe et al. (2004) and Bowerman et al. (2003) show that the number of samples we expect to obtain (i.e., a minimum of 8-12 per park in each year) should be adequate to detect a 20% increase or decrease in concentrations of most contaminants within 10 years for each park, assuming annual sampling.

For herring gulls, we will collect eggs from 13 randomly selected nests in one colony from each of the parks where colonies exist (APIS, ISRO, VOYA, and SLBE). This sampling design follows >20 years of monitoring by the Canadian Wildlife Service (CWS). Eggs will be sent to CWS for analysis of contaminants and inclusion in a larger dataset for monitoring toxics in herring gulls across the Great Lakes region. In two (SLBE and VOYA) of the four parks, a single colony exists for sampling. In one park (ISRO), one colony has been sampled by the CWS for several years and so will be included as an index site. In the remaining park (APIS), one of the two available colonies

was selected by park management because of potential disturbance to other colonial species. Because the selection of colonies at these last two parks was not random, we will analyze the data from each colony separately, through time.

Data collected from bald eagles and herring gulls will include age, sex, and physical measurements (eagles only), size and viability of eggs, presence of abnormalities in nestlings or fetuses, and location of nests and colonies. Initially, parks will be revisited every year; however, the revisit rate may be reduced after the first two pilot years, depending on data variability.

QUALITY ASSURANCE

Minimizing Sources of Error

One fundamental goal of monitoring natural resources through time is to ascertain whether a persistent, interannual, directional change is occurring in those resources within the spatial domain of interest. This hinges on the ability to measure the parameter accurately with consistent technique and adequate statistical power (i.e., sufficient sample size given variability in the indicator and desired level of confidence). A well-conceived monitoring program should identify as many of the likely primary sources of noise as possible, and envision strategies to minimize the effects of those sources of error. Broadly speaking, these sources of error fall into three categories: a) observer bias and methodological differences; b) errors in data collection, entry, and management; and c) endogenous variability in the indicator, which is not a true source of error, but is a reason that either sampling intensity or alpha must increase to maintain a given level of power.

Observer bias refers to the consistent effect that a particular observer has on values of an indicator (i.e., higher, lower, more variable, or less variable), without any actual change in the indicator itself. Observer bias can result from minor deviations in methods used, as well as from inherent differences in the ability of various observers to measure resources. To minimize effects of observer bias, we will make use of a combination of the following, depending on the nature of the indicator sampled and the sampling schedule: a) initial training and in some cases, testing, at the beginning of each field season; b) mid-season re-calibration; or c) inter-observer comparisons or explicit incorporation of observer as a covariate in analyses.

Differences in methods used are likely either to introduce bias if the correct technique is not used consistently, or increase variability if a technique is used sporadically. The value of monitoring data can be severely compromised if methods are not clearly defined and followed (Beever et al. 2005, Oakley et al. 2003). The Great Lakes Network intends to minimize the occurrence of deviations in method by adopting clearly defined protocols and standard operating procedures for each monitored indicator (see Chapter 5), following guidelines of Oakley et al. (2003). Pilot field work, in which various data collectors are given the protocols and their results compared, may be used to illustrate where the level of detail is insufficient. Because comparability with other monitoring data sets is necessary to place monitoring results within a broader (regional or national) context, the Network will seek to adopt methods that are broadly accepted as the standard method within a given discipline or taxon for the ecosystems of the region. We have collaborated with university, USGS, and other researchers and monitoring

experts in the writing and peer-review of the protocols presented in Chapter 5 to ensure that robust, widely accepted methods are employed.

The second main source of error is also human derived, and involves error in data collection, data transcription (or processing or data entry), and data management. We will address this potential source of error through a QA/QC process throughout the monitoring, as detailed in Chapters 6 and 7.

Finally, the ability to detect interannual trends in a given indicator is complicated by endogenous variability in the indicator itself (i.e., process variation). Examples of this include interannual cycles in mammals, such as the lynx-hare, moose-wolf, and microtine population cycles, and weather patterns, such as Pacific decadal oscillations, El Niño southern oscillations, and others. One approach for irruptive, cyclic, or otherwise highly variable indicators is to calculate process variability (i.e., the variance of the variance estimate over time) and the probability of conformity (that the latest observation is from the previously described distribution) (E. Rexstad, Institute of Arctic Biology and University of Alaska - Fairbanks, *personal communication*). Although this alternative approach can accommodate highly variable indicator values, it still would require longer-term data sets to obtain the same confidence in a trend than what would be required by a less variable indicator.

Strategies to Improve Effectiveness of Designs

In some cases, pilot testing may be conducted in the initial year(s) of a protocol. Reasons for conducting pilot work include documenting new methods and acquiring knowledge about park logistics, which are often difficult to ascertain without experience. One example in which pilot testing is warranted occurs in the amphibian monitoring, in which the use of parabolic reflectors is recommended to extend the area sampled, yet the method is not well documented. Duration of listening at each sampling station is also not universally agreed upon, and represents a compromise between the goals of maximizing detectability at each stop, maximizing the number of stops visited each evening, and completing surveys each night during the appropriate temporal window.

During pilot testing of any protocol, we will explore trade-offs between statistical power and Type I errors, and the value of increasing the number of visits per site versus increasing the number of sites, given restricted budgets. For certain sampling strategies (e.g., amphibians) we will adaptively refine the number of sites and re-visits made, based on an analysis of the data from pilot studies.

Using simulation analyses, Field et al. (2005) explored various aspects of allocating a limited monitoring budget to either the establishment of more sampling locations or re-visiting already established sites, to detect bird species. In cases where detectability is a known source of confounding, we will seek to employ the methods of Field et al. (2005) to make the best use of limited budget yet provide statistically powerful results and defensible interpretations.

Finally, whenever possible, we will ask quantitative ecologists from other networks to peer-review early drafts of our sampling designs in addition to collaborating with disciplinary experts during protocol development.

INTEGRATION

Integration of the various Vital Signs will occur during the design, data collection, data management, data analysis, and reporting phases of the program. This integration will occur within individual protocols, among protocols, and between this Network and other partner programs. Several of our protocols are designed to simultaneously monitor numerous variables from more than one Vital Sign (Table 4.1), such that they will be sampled at the same place (co-location) and time (co-sampling). For example, under the terrestrial vegetation protocol we expect to monitor ungulate browse, forest pests and pathogens, soils, and several metrics of forest structure, composition, and succession. Hence, forest pests and pathogens can be linked to data on forest composition and structure or soil type to provide a more holistic, integrated assessment of a given Vital Sign. Furthermore, we will integrate among protocols where possible, especially in analysis and interpretation of monitoring results. We will acquire remotely sensed data for the land cover/land use protocols to coincide temporally with data collection on terrestrial vegetation plots. The plot data will help ground-truth the remote sensing products and directly link the two data sets.

Integration will also occur between the Network's monitoring and other national and regional monitoring programs when it is scientifically valid to do so. Amphibians and landbirds, for example, will be monitored in such a way that statistically robust results are obtained for each park, yet the data are comparable with other national (e.g., NAAMP) and regional (MMP) programs. Some of these programs have accumulated > 20 years of data at > 1,000 sites around the Great Lakes, and include sites within GLKN parks. By designing protocols to collect comparable data, we will put the parks' data into a regional context, at least for a subset of response variables and spatio-temporal domains.

Similarly, water-quality monitoring for lakes and rivers will include an initial coring of bottom sediments to provide a historical record of diatom communities. Because the sensitivity and tolerance of diatoms to environmental variables – including nutrients, organic pollutants, pesticides, heavy metals, salinity (and major ion chemistry), pH, alkalinity, light, temperature, substrate, and depth – are known to vary among species (Battarbee et al. 2001), analysis of preserved diatom communities facilitates inference of past water quality. Comparing the current composition of diatom communities, which provide an integration of water quality over the short-term, to the species compositions 150-200 years ago allows us to determine whether the current conditions are within of the range of natural variability. Such information will also help the parks assess desired conditions based on the historical record.

The NPS guidelines for developing an integrated monitoring program encourage co-location of sampling sites (NPS 2003). While co-location is planned across Vital Signs, our initial protocols are not well suited to co-location because they do not exhibit spatial overlap (e.g., aquatic vs. terrestrial vs. atmospheric domains). However, sample sites selected for terrestrial vegetation and water quality, are expected to serve as 'base' sites for future monitoring protocols.

Co-location of sites has its drawbacks, however. For example, a given plot may be visited only once every five years for monitoring of terrestrial vegetation, but if co-location is forced, during the intervening years it might be visited annually by different teams to monitor breeding birds, small mammals, and deer browse. If care is not taken to

limit the effects of each monitoring team's visit, the monitoring could show change in the vegetation due solely to the disturbance imposed by the sampling teams (*sensu* Paquin 2004 and Eckrich and Holmquist 2000). Additionally, co-location assumes that the same points are equally valid to sample the various target domains for each of the monitoring programs – an assumption that may not always hold. For the above reasons, we did not force co-location for protocols; however, by developing key protocols first, we increase the likelihood for co-location, if appropriate. Those developing new protocols will have, as their first option, a set of probabilistically chosen sites or plots to use. The choice of whether to adopt these sites will depend on: a) whether it is ecologically appropriate for the metrics being monitored, b) whether it is statistically appropriate (in terms of sample size and spatial allocation), and c) whether it will affect the quality of other data being collected at those locations.

In addition to integration in the field, we will integrate data analytically. The conceptual models that provide the linkage among Vital Signs (Gucciardo et al. 2004) were based on known or proposed linkages among factors that operate across spatial and temporal scales. Given the data collected across protocols, we can assess the presence and strength of these relationships using a diversity of statistical techniques, ranging from simple correlations to structural equation models. It must be noted, however, that the primary goal of the protocols was to develop statistically sound monitoring for long-term change detection; tests of causality would require a very different sampling design. That stated, it is still feasible to use GIS-based analyses, simple linear models, and more advanced techniques such as multivariate analyses (e.g., canonical correspondence, redundancy analyses, and classification and regression trees (CART); McCune and Grace 2002), structural equation modeling (SEM), and Bayesian approaches to quantify relationships noted in the GLKN conceptual models. These statistical approaches are described more fully in Chapter 7.